Theses of the PhD thesis

Improvements of Silent Speech Interface Algorithms

Amin Honarmandi Shandiz

Supervisor:
László Tóth, PhD, Associate Professor

Doctoral School of Computer Science
University of Szeged

Department of Computer Algorithms and Artificial Intelligence

2023

1 Introduction

This PhD thesis focuses on advancing the field of Silent Speech Interface, by proposing new strategies and techniques to improve various aspects of the ultrasound SSI project. The project has wide range of usability including when humans can not talk due to a specific disease related to speech or they don't want to talk due to several reasons. In this thesis the goal is to enhance model implementation, data preparation and processing, generalization, and model training speed. The methods proposed in this thesis have been tested on two large datasets, and the results have been thoroughly evaluated. In this introduction, we provide an overview of the SSI system's fundamental components, including the feature extractor, different modalities, and model training. We also explore the application of deep neural networks, particularly Convolutional Neural Networks (CNNs), and their effectiveness in image processing. As it is shown in Figure 1, speech could be synthesize from articulatory signals which are related to human's speech. [13].

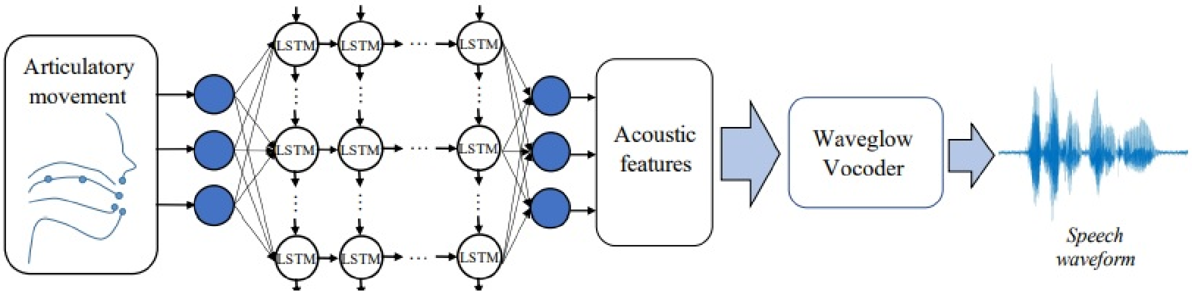


Figure 1: *Process of transforming articulatory features to acoustic features using deep learning methods for speech synthesis[13]*

The subsequent chapters delve into specific aspects of the SSI system and adapt algorithms from related areas to enhance their performance.

2 3D Convolutional Neural Networks for Developing Silent Speech Interfaces Utilizing Ultrasound

Chapter 2 of this thesis investigates the use of deep neural networks for converting ultrasound videos of tongue movements into speech. Specifically, Convolutional Neural Networks (CNNs) are employed to process a sequence of images, a technique widely recognized for image recognition tasks. The input to the CNNs is a video sequence that captures the temporal trajectory of tongue movements. The chapter explores different network structures for processing time sequences, including the stacking of a 2D CNN and a recurrent neural network (RNN), as well as extending the 2D CNN to a 3D CNN by incorporating time as an additional dimension [14, 19, 21, 22, 24, 27, 33]. The experimental results reveal that the 3D CNN model achieves a lower error rate, requires a smaller model size, and trains faster compared to the CNN+LSTM model. This finding suggests that 3D CNNs offer a viable alternative to recurrent neural models for ultrasound video-based SSI systems. The application of CNNs for processing time sequences presents a novel approach that achieves superior performance with faster training times, holding promise for improving silent speech technology.

3 Utilizing adversarial training to improve Deep Neural Network models

Chapter 3 aimed to enhance the performance of SSI models by leveraging Generative Adversarial Networks (GANs) and incorporating perceptual loss in addition to conventional loss functions.

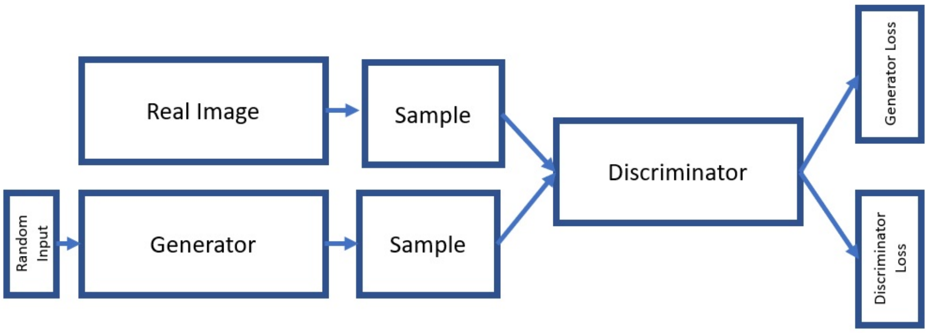
In deep neural network based ultrasound image SSI projects, conventional loss functions for 3D CNNs (3D convolutional neural networks) are often used to optimize the performance of the models. These loss functions can be simple mathematical formulas that measure the difference between the predicted output of the model and the actual ground truth. The most commonly used loss function for image processing tasks is the mean squared error. However, conventional loss functions for 3D CNNs in ultrasound image SSI projects face several challenges. For example, conventional loss functions do not always capture the perceptual quality of the output spectrogram images. Additionally, conventional loss functions may result in blurry spectrogram images or spectrogram images with loss of fine details. Therefore, there is a need to use alternative loss functions that can capture the perceptual quality of the spectrogram images. Perceptual loss functions are one type of alternative loss functions that can improve the performance of ultrasound image SSI models. Perceptual loss functions are originally designed to measure the similarity between two images in terms of their perceptual qualities, such as texture, contrast, and sharpness, instead of just measuring the pixel-wise differences between the images. Perceptual loss functions are commonly used in image style transfer and image generation tasks [17, 20, 25].

The objective of this chapter was to generate high-quality spectrogram images with improved accuracy and fidelity, crucial for the SSI project. The proposed method involved introducing GAN as a perceptual loss term to the conventional loss function employed in the SSI model. This perceptual loss term quantified the difference between features extracted from real and synthetic predicted output using a pre-trained neural network. By doing so, the GAN could generate output that not only matched the target distribution but also captured the relevant features and structures of real data. The performance of the proposed method was evaluated using two distinct datasets: one comprising Hungarian sentences and the other containing English sentences.

3.1 Generative Adversarial Networks

A Generative Adversarial Network (GAN) is a type of neural network that consists of two main components: a generator network and a discriminator network. The generator network is responsible for generating synthetic data, while the discriminator network's job is to distinguish between synthetic data and real data [15]. These two networks are trained together in a process called adversarial training, where the generator tries to produce more realistic data, and the discriminator tries to improve its ability to differentiate between real and fake data. In the GAN structure, the generator network takes a random noise vector

Figure 2: *A typical GAN implementation consists of two neural networks (generator and discriminator) trained in an adversarial manner, with noise vectors as input to the generator and real/fake labels as input to the discriminator.*



as input and generates a synthetic image(here spectrogram) as output. The discriminator network then takes the synthetic image generated by the generator network and a real image from the dataset as input. It outputs a probability value indicating whether the input image is real or fake. The training process is adversarial, where the generator aims to produce more realistic images to deceive the discriminator, while the discriminator aims to improve its ability to classify the images accurately.

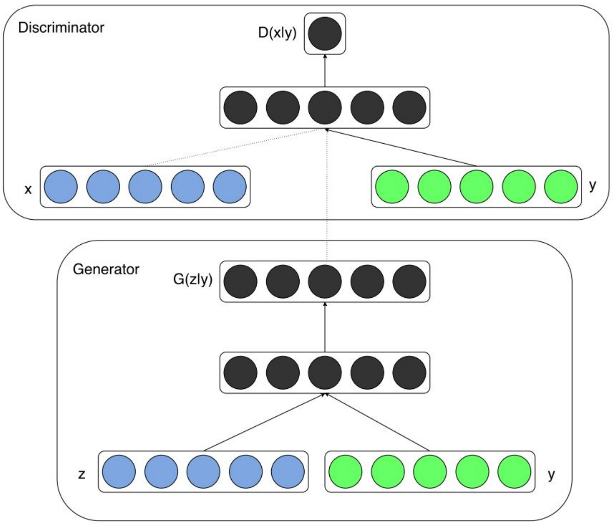
Through this adversarial training process, the generator learns to generate more realistic images, and the discriminator learns to accurately classify the images as real or fake. Once the training is complete, the generator can be used to generate new images that resemble the real images in the dataset. GANs have been successfully applied in various domains such as image generation, speech synthesis, and natural language processing.

3.2 Conditional Generative Adversarial Networks

Conditional Generative Adversarial Networks (cGANs) are a type of Generative Adversarial Network (GAN) that incorporate additional information, known as conditions, into the generator and discriminator models [64]. These conditions can take the form of labels, attributes, or any other relevant data that helps guide the generation process. By including conditions(see fig 3), cGANs provide more precise control over the generated output. cGANs have been successfully applied to various tasks, including image-to-image translation, text-to-image generation, and style transfer [26]. Image-to-image translation involves converting an image from one domain to another, such as transforming a grayscale image into a colored image or turning a sketch into a realistic image. Text-to-image generation aims to generate an image based on a textual description. Style transfer involves transferring the style of one image onto another while preserving the content, enabling the creation of new stylized images.

The key advantage of cGANs is their ability to incorporate conditional information, which allows for targeted and controlled generation. This makes them highly applicable in fields such as computer vision, natural language processing, and creative arts. They offer opportunities for more precise image generation, enabling tasks that require specific attributes or characteristics to be fulfilled.

Figure 3: *Conditional Generative Adversarial Networks*



3.3 Generative Adversarial Networks for Articulatoryto-Acoustic Mapping

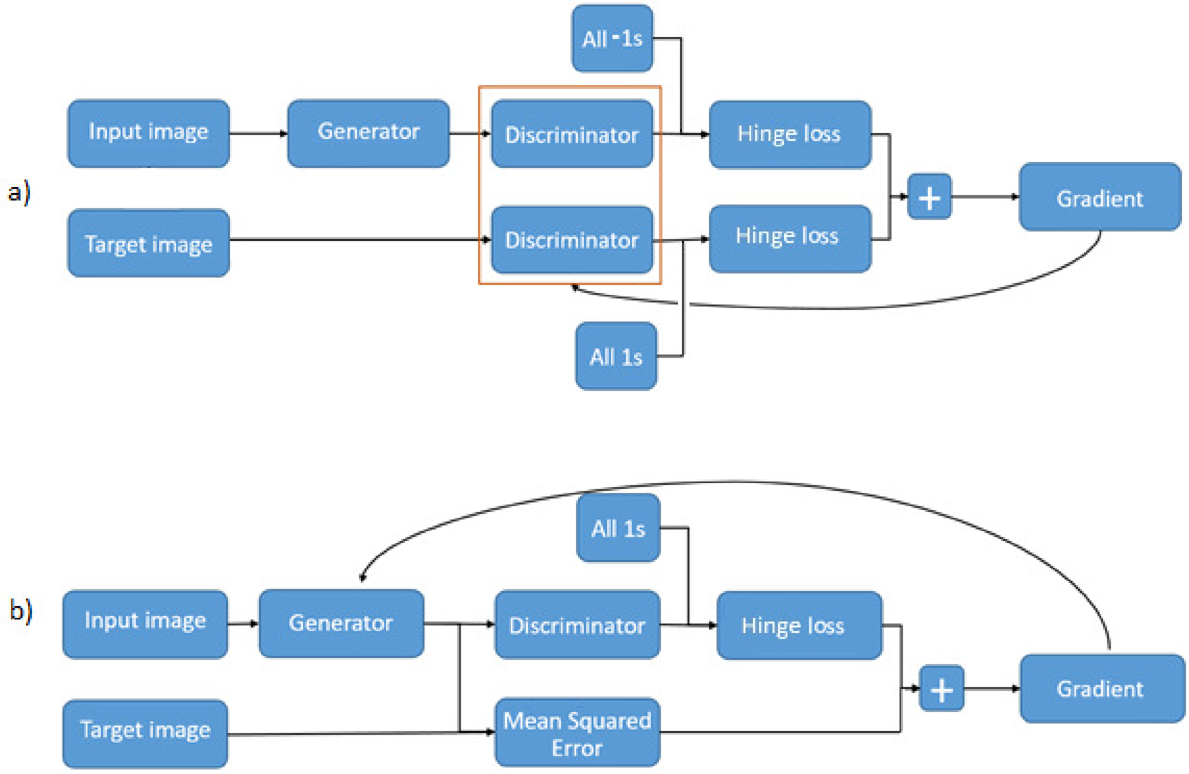
In our case, the generator network was responsible for converting ultrasound data to mel-spectral data, using the same network that showed the better result in chapter **??**. The generator and discriminator were trained in parallel using a two-step process. In the first step, the discriminator was trained on real and generated spectrograms using the hinge loss function. In the second step, the generator weights were updated using a combination of the mean squared error (MSE) loss and the discriminator's feedback. This adversarial training approach aimed to create generator outputs that resemble real spectrograms.

The results demonstrated that incorporating the perceptual loss led to a significant improvement in spectral quality and accuracy. This approach has the potential to enhance the reliability and accuracy of SSI models, thereby advancing SSI applications. Moreover, the utilization of GANs and perceptual loss can be extended to other modalities, opening up possibilities for further advancements in the field.

4 Neural Speaker Embeddings for Generalizing Ultrasound SSI model

In Chapter 4, we focused on enhancing the SSI model by exploring Embedding Neural Networks. The previous performance of the SSI model was hindered by suboptimal parameter tuning in unseen speakers. To overcome this issue, we introduced a novel approach called x-vector, which aimed to improve the model's performance by incorporating speaker information into the input ultrasound data [31]. The effectiveness of this approach was evaluated on unseen speakers, and its impact on improving the model's generalizability was assessed. The evaluation was conducted using an English dataset prepared at both the frame and speaker levels, including their respective spectrograms. The x-vector concept was introduced as a neural solution to replace the Gaussian-based i-vector approach

Figure 4: *The error calculation (forward arrows) and weight update (backward arrows) training steps for the discriminator (upper image) and the generator (lower image) networks of the GAN.*



for speaker recognition [29]. It consists of a deep neural network (DNN) trained to discriminate speakers, with three main parts. The lower layers, typically a time-delay network (TDNN), operate on frame-level information. The subsequent temporal pooling layer aggregates statistics over speech segments, and the aggregated values are processed by fully connected layers. The network produces a fixed-size speaker embedding vector. In the context of ultrasound-based silent speech interfaces, the x-vector DNN was adjusted to operate with ultrasound images, with a frame-level part utilizing a 3D convolutional layer and a statistical pooling layer performing simple average pooling. The segment-level part consists of fully connected layers and a softmax output layer. The results demonstrated that the integration of the x-vector approach significantly enhanced the model's ability to generalize and perform better on unseen speakers.

5 Convolutional Neural Networks for Detecting Voice Activity in Silent Speech Interfaces based on Ultrasound

In Chapter 5, we investigated the utilization of ultrasound images for differentiating between silent and speech segments, (Figure 6), similar to voice activity detection in speech.

Figure 5: *Illustration of the UTI-based xx-vector network.*

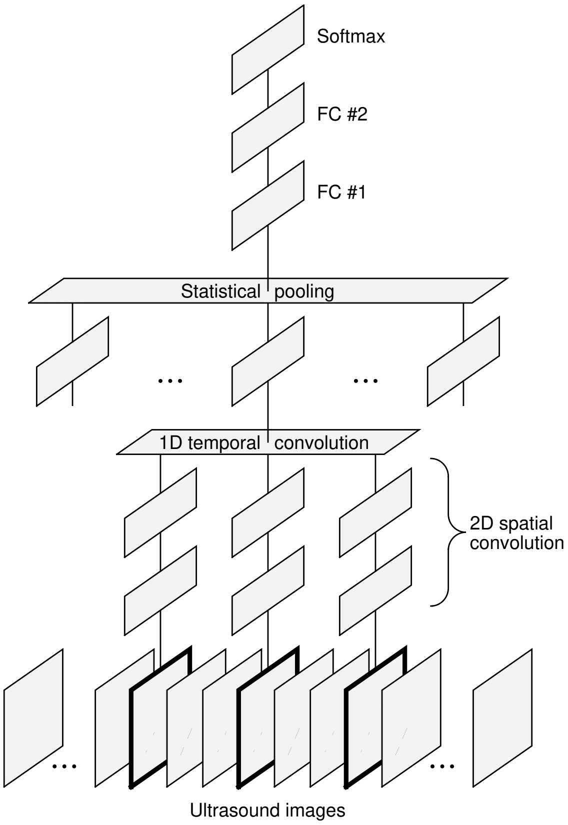


Fig 6 shows two examples of the tongue position recorded by the device, when the subject is speaking (producing a vowel) and when he is not – the diagonal light stripes in the images correspond to the tongue of the speaker. After examining several samples, we got the impression that speaking versus remaining silent typically results in more drastic changes in the speech signal than in the corresponding ultrasound tongue images, so voice activity detection based on the latter is presumably much harder. In the following we train a CNN to perform the voiced/invoiced classification using such ultrasound images. The structure of this VAD-CNN and the network that we apply for the SSI task are very similar.

First, we estimated training labels based on a public Voice Activity Detection (VAD) implementation applied to parallel speech recordings [32]. The classifier developed for this purpose achieved an accuracy of 86% in discriminating silence and speech frames, Figure 7. Furthermore, we emphasized the importance of carefully handling the amount of silence retained in the training set, as an excessive presence of silence can adversely affect the quality of the synthesized speech and the training process of the model.

Figure 6: *Two UTI examples from the database, one for a speech (vowel) frame (left) and one for a silent frame (right).*

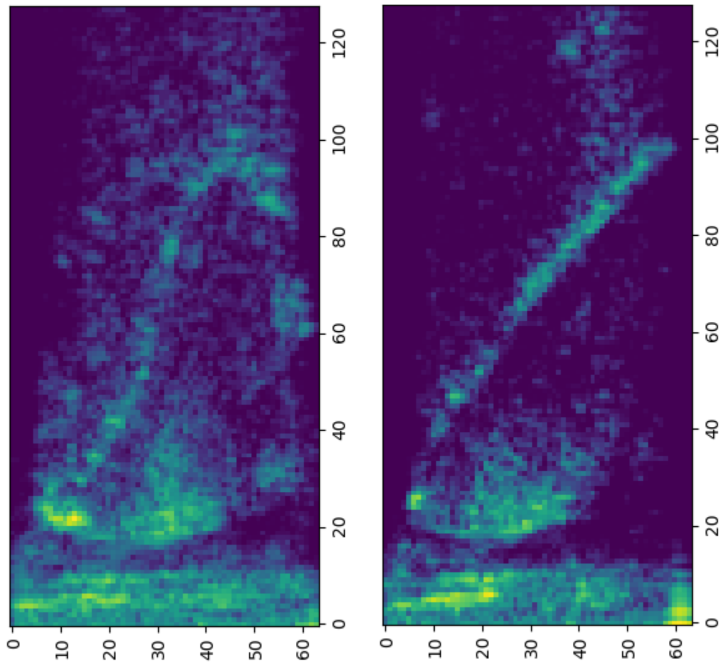
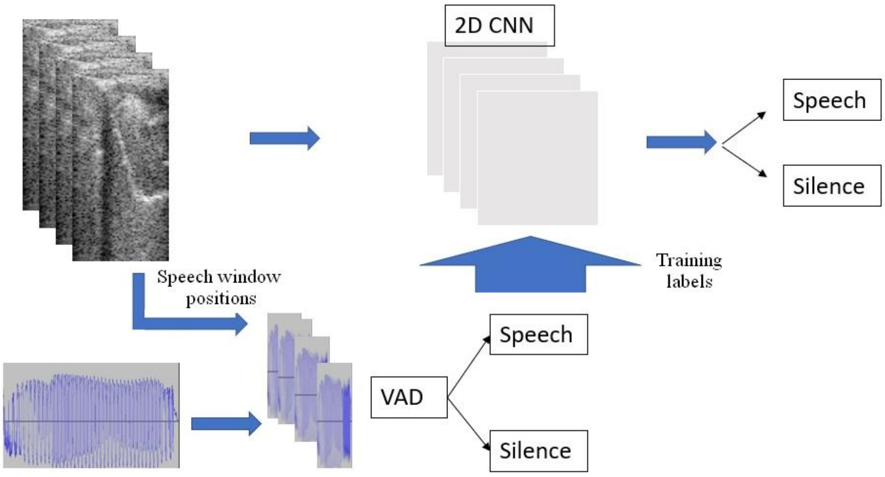


Figure 7: *Illustration of obtaining the VAD training labels and training the 2D-CNN for silence/speech classification.*

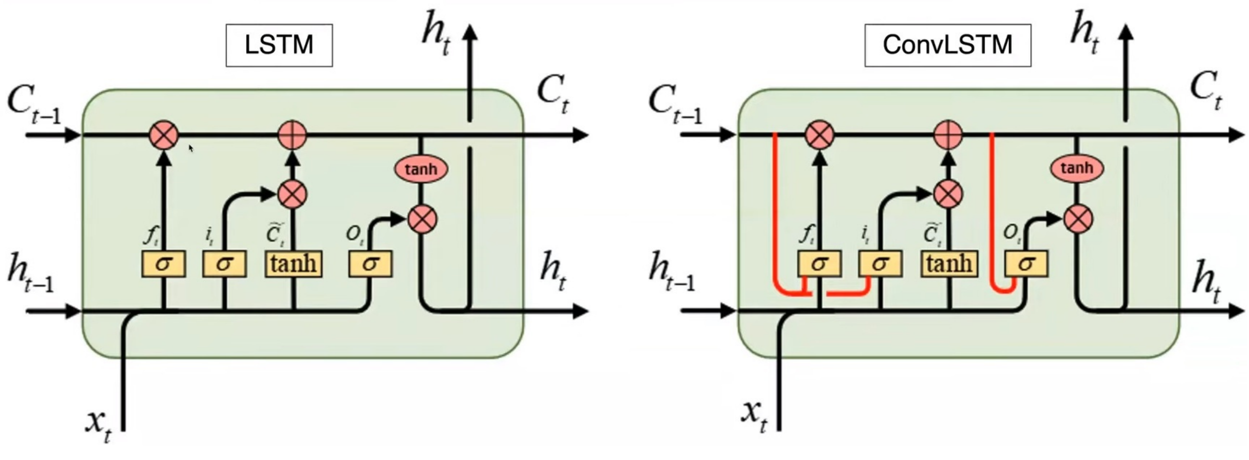


To address the challenges associated with preprocessing in Silent Speech Interfaces, we proposed a novel method of voice activity detection. Instead of considering the entire speech, we implemented VAD on each ultrasound frame, resulting in better alignment between the spectrogram and ultrasound frames, and consequently, more reliable features for the model. We employed the VAD technique to remove silence as a preliminary step before feature extraction. Subsequently, we synchronized the windowed speech signal with the ultrasound frames and fed them into the VAD process. The retained frames determined by VAD were then subjected to subsequent feature extraction steps, which facilitated tasks such as speech synthesis. By incorporating VAD at the ultrasound frame level, rather than at the speech level, we observed improved accuracy in the SSI model. The proposed VAD method effectively removed silence from the input data and yielded more reliable features, demonstrating the potential of ultrasound images for discriminating between silent and speech segments. This finding holds significant implications for the advancement of the SSI technology.

6 Enhanced analysis of ultrasound tongue videos via the fusion of ConvLSTM and 3D Convolutional Networks

In Chapter 6, we addressed the challenge of high computational cost associated with deep learning models, which typically require large amounts of data to achieve optimal performance. Our proposed solution involved a fusion of Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), known as ConvLSTM [30],which was applied in another areas such as emotion recognition [23] and in [34] with promising results applied along with the integration of a 3D Convolutional Network (Conv3D), (Figure 8). This fusion enabled the extraction of both sequential and volumetric information from the data while reducing the number of layers and parameters required for training. Consequently, our approach achieved high performance with improved efficiency.

Figure 8: *Internal structure of a standard LSTM cell and its extended version (with extra peephole connections) used in Convolutional LSTMs [11, 12].*



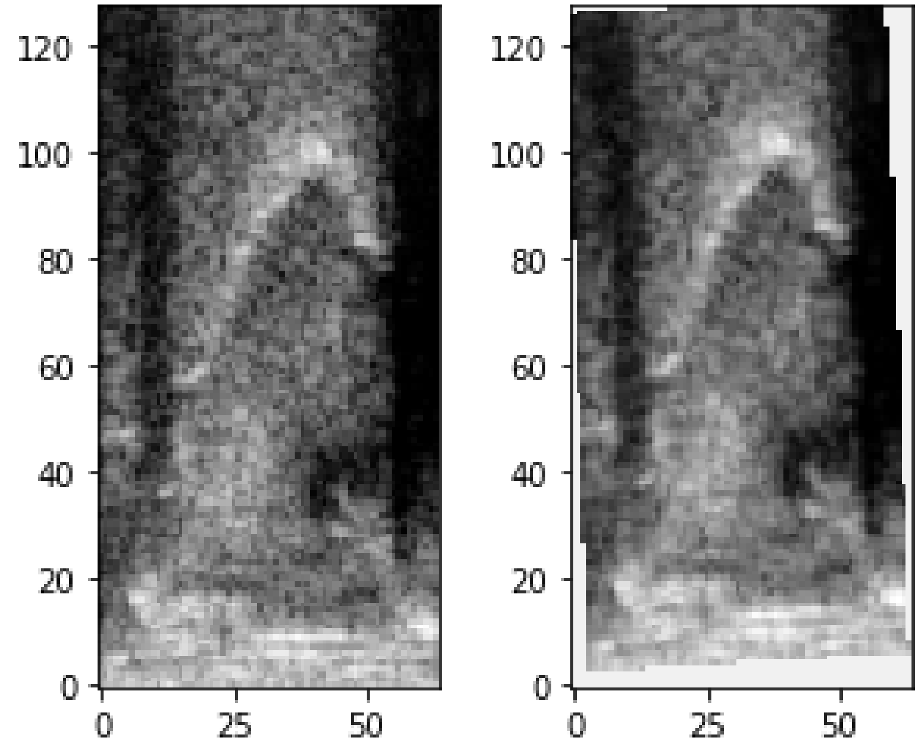
To evaluate the effectiveness of our proposed method, we conducted experiments using a Hungarian dataset and compared the results against previous state-of-the-art models. This chapter provides a detailed explanation of the ConvLSTM model's implementation and showcases its ability to extract spatial and temporal features from ultrasound tongue videos. The experimental results demonstrated that our proposed method outperformed the previous state-of-the-art models, delivering superior accuracy and efficiency.

In conclusion, our method offers a promising approach to enhance the performance of deep learning models while mitigating the computational burden. The potential impact of our approach extends to the field of image processing and deep learning, where it can contribute to advancements in research and application.

7 Enhancing Tongue Ultrasound-Based Silent Speech Interfaces with Spatial Transformer Networks

Chapter 7 explores the application of spatial transformer networks (STNs) to enhance the speaker and session adaptation capabilities of ultrasound tongue imaging-based silent speech interfaces. Traditional SSI models often exhibit limited performance when transitioning between different speakers or sessions due to their speaker-specific nature [18]. In this area, There have already been several cross-session and cross-speaker studies. To handle the session dependency of UTI based synthesis, by using data from different sessions [16], unsupervised model adaption [28], or as we used in chapter 4 by using xx-vectors featurs from the speakers.

Figure 9: *An example UTI image, before and after STN.*



Most of the above approaches hope to solve speaker sensitivity simply by acquiring articulatory training data from a large quantity of speakers. In this chapter, we experiment with a direct adaptation of an UTI-based SSI network to the actual speaker or session.

To address this, we proposed an augmenting deep networks with a spatial transformer network module. The STN module facilitates affine transformations on input images, enabling quick adaptation for different speakers and sessions. By integrating spatial transformer networks into the SSI models, we can improve their performance when confronted with variations in tongue articulation across speakers or changes in the recording setup. This enhancement enhances the overall flexibility and adaptability of ultrasound tongue-based SSIs, leading to improved synthesis performance and a broader range of applicability.

By investigating the integration of STNs in Chapter 7, we aim to overcome the limitations associated with speaker and session adaptation in SSI models, ultimately advancing the field of silent speech interfaces based on ultrasound tongue imaging. When only the STN module was adapted, the error rate decreased significantly. When both the STN module and the linear output layer were allowed to adapt, the error reduction went even more. Although the improvement was slightly smaller for 3D input blocks, similar tendencies were observed.

8 Contributions of the thesis

In the **first thesis group**, the contributions are related to the publication ’3D Convolutional Neural Networks for Developing Silent Speech Interfaces Utilizing Ultrasound’. Detailed discussion can be found in Chapter 2.

I/1. Implementing the neural networks used in the experiments to restore speech signals from articulatory recordings. Specifically, we implemented a 3D convolutional neural network with different window length and compared it with different variations of CNN and combination with CNN+LSTM and BiLSTM networks.

I/2. Calculating the performance of the models using objective metrics such as STOI, PESQ, and MCD. The results obtained from these metrics were analyzed to compare the performance of the different network architectures.

In the **second thesis group**, the contributions are related to the publication ’Utilizing adversarial training to improve Deep Neural Network models’. Detailed discussion can be found in Chapter 3.

II/1. Implementation of GAN and CGAN models for image generation.

II/2. Utilization of various SSI models as generators in the GAN and CGAN frameworks.

II/3. Calculation and analysis of performance metrics to evaluate the effectiveness of the models.

II/4. Conducting research in the field of GANs and image generation.

In the **third thesis group**, the contributions are related to the publication ’Neural Speaker Embeddings for Generalizing Ultrasound SSI model’. Detailed discussion can be found in Chapter 4.

III/1. Preparing data for the experiments.

III/2. Implementing the model and conducting the experiments.

III/3. Comparing the results obtained from the experiments.

III/4. Calculating the relevant metrics to evaluate the model's performance.

In the **forth thesis group**, the contributions are related to the publication 'Convolutional Neural Networks for Detecting Voice Activity in Silent Speech Interfaces based on Ultrasound'. Detailed discussion can be found in Chapter 5.

IV/1. Implementing the model for voice activity detection using a CNN architecture and training it with binary cross-entropy loss function.

IV/2. Developing the idea of applying VAD to remove silence from the speech signal in SSI systems.

IV/3. Analyzing the results of experiments with different amounts of silence in the corpus, comparing the MCD and MSE metrics, and evaluating the impact of VAD on the SSI.

In the **fifth thesis group**, the contributions are related to the publication 'Enhanced analysis of ultrasound tongue videos via the fusion of ConvLSTM and 3D Convolutional Networks'. Detailed discussion can be found in Chapter 6.

V/1. Preparing data for the specific task.

V/2. Implementing code for the models (Conv3D, Conv3D+BiLSTM, ConvLSTM).

V/3. Analyzing and interpreting the results.

V/4. Calculating the evaluation metric for the results.

In the **six thesis group**, the contributions are related to the publication 'Enhancing Tongue Ultrasound-Based Silent Speech Interfaces with Spatial Transformer Networks'. Detailed discussion can be found in Chapter 7.

VI/1. Preparing data.

VI/2. Implementing code for the models.

VI/3. Analyzing and interpreting the results.

Table 1 summarizes the relation between the thesis points and the corresponding publications.

Table 1: *Correspondence between the thesis points and my publications.*

Publication	**Thesis point**						
	II/1	II/2	III/1	IV/1	V/1	VI/1	VII/1
[2]	●						
[3]			●				
[4]				●			
[5]					●		
[6]						●	
[9]							●
[10]		●					

The author's publications on the subjects of the thesis

Journal publications

[1] T. G. Csapó, G. Gosztolya, L. Tóth, **A. Honarmandi Shandiz,** , A. Markó. Optimizing the Ultrasound Tongue Image Representation for Residual Network-based Articulatory-to-Acoustic Mapping. *Sensors*, 22(22), 8601, 2022.

Full papers in conference proceedings

[2] L. Toth, **A. Honarmandi Shandiz**. 3D convolutional neural networks for ultrasound-based silent speech interfaces. In *International Conference on Artificial Intelligence and Soft Computing*, 159-169, Springer, 2020.

[3] **A. Honarmandi Shandiz**, T. G. Csapó, G. Gosztolya, L. Tóth, A. Markó. Improving neural silent speech interface models by adversarial training. In *Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV)*, 430-440, Springer, 2021.

[4] **A. Honarmandi Shandiz**, L. Tóth, G. Gosztolya, A. Markó, T. G. Csapó. Neural Speaker Embeddings for Ultrasound-Based Silent Speech Interfaces. In *Proceedings of the International Conference on Interspeech*, 1932-1936, Springer, 2021.

[5] **A. Honarmandi Shandiz**, L. Tóth. Voice activity detection for ultrasound-based silent speech interfaces using convolutional neural networks. In *Text, Speech, and Dialogue: 24th International Conference*, 499-510, Springer, 2021.

[6] **A. Honarmandi Shandiz**, L. Tóth. Improved Processing of Ultrasound Tongue Videos by Combining ConvLSTM and 3D Convolutional Networks. In *Advances and Trends in Artificial Intelligence. Theory and Practices in Artificial Intelligence: 35th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, 265–274, Springer, 2022.

[7] Y. Yide, **A. Honarmandi Shandiz**, L. Tóth. Reconstructing speech from real-time articulatory MRI using neural vocoders. In *2021 29th European Signal Processing Conference (EUSIPCO)*, 245–249, IEEE, 2021.

[8] C. Zainkó, L. Tóth, **A. Honarmandi Shandiz**, G. Gosztolya, A. Markó, G. Németh, T. G. Csapó. Adaptation of Tacotron2-based Text-To-Speech for Articulatory-to-Acoustic Mapping using Ultrasound Tongue Imaging. In *11th ISCA Speech Synthesis Workshop (SSW 11)*, 54-59, Springer, 2021.

[9] L. Tóth, **A. Honarmandi Shandiz**, G. Gosztolya, T. G. Csapó. Adaptation of Tongue Ultrasound-Based Silent Speech Interfaces
Using Spatial Transformer Networks. In *Proceedings of the International Conference on Interspeech*, Springer, 2023.

Further related publications

[10] T. G. Csapó, L. Tóth, **A. Honarmandi Shandiz**,G. Gosztolya, A. Markó. 3D konvolúciós neuronhálón és neurális vokóderen alapuló némabeszéd-interfész. In *MSZNY*, 2021.

Other References

[11] Convolutional LSTM. https://medium.com/neuronio/an-introduction-to-convlstm-55c9025563a7, 2019.

[12] Recurrent neural networks and LSTMs with keras. https://blog.eduonix.com/artificial-intelligence/recurrent-neural-networks-lstms-keras, 2020.

[13] B. Cao, A. Wisler, and J. Wang. Speaker adaptation on articulation and acoustics for articulation-to-speech synthesis. *Sensors*, 22(16):6056, 2022.

[14] Jose A. Gonzalez, Lam A. Cheah, Angel M. Gomez, Phil D. Green, James M. Gilbert, Stephen R. Ell, Roger K. Moore, and Ed Holdsworth. Direct speech reconstruction from articulatory sensor data by machine learning. *IEEE/ACM Trans. ASLP*, 25(12):2362–2374, 2017.

[15] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

[16] G. Gosztolya, T. Grósz, L. Tóth, A. Markó, and T. G. Csapó. Applying DNN Adaptation to Reduce the Session Dependency of Ultrasound Tongue Imaging-Based Silent Speech Interfaces. *Acta Polytechnica Hungarica*, 17(7):109–124, 2020.

[17] ITU-R. ITU-R recommendation BS.1534: Method for the subjective assessment of intermediate audio quality, 2001.

[18] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu. Spatial transformer networks. In *NIPS 28*, pages 2017–2025. 2015.

[19] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(1):221–231, 2013.

[20] L. R. Jonathan, W. Scott, H Erdogan, and J. R. Hershey. SDR - half-baked or well done. In *Proc. ICASSP*, 2019.

[21] M. Kim, B. Cao, T. Mau, and J. Wang. Speaker-Independent Silent Speech Recognition From Flesh-Point Articulatory Movements Using an LSTM Neural Network. *IEEE/ACM Trans. ASLP*, 25(12):2323–2336, 2017.

[22] N. Kimura, M. Kono, and J. Rekimoto. Sottovoce: An ultrasound imaging-based silent speech interaction using deep neural networks. In *Proc. of CHI Conf. on Human Factors in Computing Systems*, 2019.

[23] S. Kwon et al. CLSTM: Deep feature-based speech emotion recognition using the hierarchical ConvLSTM network. *Mathematics*, 8(12):2133, 2020.

[24] Z. C. Liu, Z. H. Ling, and L. R. Dai. Articulatory-to-acoustic conversion using BLSTM-RNNs with augmented input representation. *Speech Communication*, 99(2017):161–172, 2018.

[25] J. Martín-Doñas, A. Gomez, J. Gonzalez Lopez, and A. Peinado. A deep learning loss function based on the perceptual evaluation of the speech quality. *IEEE Signal Processing Letters*, 25(11):1680 – 1684, 2018.

[26] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[27] E. Moliner and T. Csapó. Ultrasound-based silent speech interface using convolutional and recurrent neural networks. *Acta Acustica united with Acustica*, 105, 2019.

[28] M. Sam Ribeiro, A. Eshky, K. Richmond, and S. Renals. Silent versus modal multi-speaker speech recognition from ultrasound and video. In *Proc. Interspeech*, 2021.

[29] A Senior and I. Lopez-Moreno. Improving DNN speaker-independence with I-vector inputs. In *Proc. ICASSP*, pages 225–229, 2014.

[30] X. Shi, Z. Chen, H. Wang, D. Yeung, W. Wong, and W. Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *arXiv preprint arXiv:1506.04214*, 2015.

[31] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur. X-vectors: Robust DNN embeddings for speaker recognition. In *Proc. ICASSP*, pages 5329–5333, 2018.

[32] E. Verteletskaya and K. Sakhnov. Voice activity detection for speech enhancement applications. *Acta Polytechnica*, 50(4), 2010.

[33] C. Wu, S. Chen, G. Sheng, P. Roussel, and B. Denby. Predicting tongue motion in unlabeled ultrasound video using 3D convolutional neural networks. In *Proc. ICASSP*, pages 5764–5768, 2018.

[34] C. Zhao, P. Zhang, J. Zhu, C. Wu, H. Wang, and K. Xu. Predicting tongue motion in unlabeled ultrasound videos using convolutional LSTM neural networks. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5926–5930. IEEE, 2019.

9 Összefoglalás

Ez a doktori értekezés a Néma Beszéd Interface (SSI) területének előmozdítására összpontosít, új stratégiák és technikák javaslatával a hangtompított ultrahang SSI projekt különböző szempontjainak fejlesztése érdekében. A projekt széles körű használhatóságot kínál, ideértve azokat az eseteket, amikor az emberek nem tudnak beszélni a beszédhez kapcsolódó konkrét betegség miatt, vagy több okból nem szeretnének beszélni. Ebben az értekezésben a cél a modell implementáció, adat előkészítés és feldolgozás, általánosítás és modellképzés sebességének javítása. Az értekezésben javasolt módszereket két nagy adathalmazon tesztelték, és az eredményeket alaposan értékelték. Az bevezetésben áttekintést adunk az SSI rendszer alapvető összetevőiről, beleértve a jellemzők kinyerését, különböző modalitásokat és a modellképzést. Emellett felfedezzük a mély neurális hálózatok, különösen a Konvolúciós Neurális Hálózatok (CNN-ek) alkalmazását és hatékonyságukat képfeldolgozásban .

Declaration

In the PhD dissertation of Amin Honarmandi Shandiz entitled **“Improvements of Silent Speech Interface Algorithms”**, with list of publications :

[1] L. Toth, **A. Honarmandi Shandiz**. 3D convolutional neural networks for ultrasound-based silent speech interfaces. In *International Conference on Artificial Intelligence and Soft Computing*, 159-169, Springer, 2020.

[2] **A. Honarmandi Shandiz**, T. G. Csapó, G. Gosztolya, L. Tóth, A. Markó. Improving neural silent speech interface models by adversarial training. In *Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV)*, 430-440, Springer, 2021.

[3] **A. Honarmandi Shandiz**, L. Tóth, G. Gosztolya, A. Markó, T. G. Csapó. Neural Speaker Embeddings for Ultrasound-Based Silent Speech Interfaces. In *Proceedings of the International Conference on Interspeech*, 1932-1936, Springer, 2021.

[4] **A. Honarmandi Shandiz**, L. Tóth. Voice activity detection for ultrasound-based silent speech interfaces using convolutional neural networks. In *Text, Speech, and Dialogue: 24th International Conference*, 499-510, Springer, 2021.

[5] **A. Honarmandi Shandiz**, L. Tóth. Improved Processing of Ultrasound Tongue Videos by Combining ConvLSTM and 3D Convolutional Networks. In *Advances and Trends in Artificial Intelligence. Theory and Practices in Artificial Intelligence: 35th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, 265–274, Springer, 2022.

[6] L. Tóth, **A. Honarmandi Shandiz**, G. Gosztolya, T. G. Csapó. Adaptation of Tongue Ultrasound-Based Silent Speech Interfaces
Using Spatial Transformer Networks. In *Proceedings of the International Conference on Interspeech*, Springer, 2023.

Amin Honarmandi Shandiz contribution was decisive in the following results:

- In **Thesis II**, the focus is on implementing and developing a 3D CNN. Different combinations with LSTM and Bi-LSTM networks are experimented with, and the model's performance is evaluated using objective metrics[1].
- In **Thesis III**, involves experimenting with various variations of Generative Adversarial Networks (GAN) and Conditional GAN (CGAN). These models are implemented as generators, and performance metrics are calculated based on the generated results [2].
- In **Thesis IV**, involves preparing data for this specific section. Experiments and results are conducted using models that utilize different extracted embedding features as x-vectors. The performance of these models is evaluated [3].

- In **Thesis V**, focuses on data preparation for a particular experiment. The Voice Activity Detection (VAD) model is implemented, and the experiment involves testing the model with a new preprocessing strategy. The duration of silence within the data is varied to compare its influence on the model's results. Performance metrics are calculated [4].

- In **Thesis VI**, explores the use of the ConvLSTM model for SSI. This method is considered with combination with other networks such as LSTM and 3D CNN. The results are evaluated to assess the effectiveness of the approach[5].

- In **Thesis VII**, analyzes the implementation of the method on both 2D and 3D data using a 2D Spatial Transformer Network (STN). Additionally, a 3D STN is implemented and tested on the data. The results are analyzed, the model is evaluated, and the outcomes are visualized[6].

these results cannot be used to obtain an academic research degree, other than the submitted PhD thesis of Amin Honarmandi Shandiz.

Szeged, 2023.07.11

Amin Honarmandi Shandiz
PhD condidate

Lászlo Tóth, PhD.
Supervisor

The head of the Doctoral School of Computer Science declares that the declaration above was sent to all of the coauthors and none of them raised any objections against it.

Szeged, 2023.07.

SZEGEDI TUDOMÁNYEGYETEM
INFORMATIKA DOKTORI ISKOLA
SZEGED

Mark Jelasity, DSc
Head of Doctoral School